Received: 23 August 2021 | Revised: 2 March 2022 | Accepted: 21 March 2022

DOI: 10.1111/1755-0998.13611

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Machine learning models identify gene predictors of waggle dance behaviour in honeybees

Marcell Veiner[1] | Juliano Morimoto[2] | Ellouise Leadbeater[3] | Fabio Manfredini[2,3]

[1]The School of Natural and Computing Sciences, University of Aberdeen, Aberdeen, UK

[2]The School of Biological Sciences, University of Aberdeen, Aberdeen, UK

[3]School of Biological Sciences, Royal Holloway University of London, Egham, UK

Correspondence
Fabio Manfredini, The School of Biological Sciences, University of Aberdeen, Aberdeen Scotland, UK.
Email: fabio.manfredini@abdn.ac.uk

Marcell Veiner, The School of Natural and Computing Sciences, University of Aberdeen, Aberdeen Scotland, UK.
Email: veinermarcell@gmail.com

Funding information
Natural Environment Research Council, Grant/Award Number: NE/S007377/1; H2020 European Research Council, Grant/Award Number: 638873

Handling Editor: Nick Fountain-Jones

Abstract

The molecular characterization of complex behaviours is a challenging task as a range of different factors are often involved to produce the observed phenotype. An established approach is to look at the overall levels of expression of brain genes—or 'neurogenomics'—to select the best candidates that associate with patterns of interest. However, traditional neurogenomic analyses have some well-known limitations: above all, the usually limited number of biological replicates compared to the number of genes tested—known as the "curse of dimensionality." In this study we implemented a machine learning (ML) approach that can be used as a complement to more established methods of transcriptomic analyses. We tested three supervised learning algorithms (Random Forests, Lasso and Elastic net Regularized Generalized Linear Model, and Support Vector Machine) for their performance in the characterization of transcriptomic patterns and identification of genes associated with honeybee waggle dance. We then matched the results of these analyses with traditional outputs of differential gene expression analyses and identified two promising candidates for the neural regulation of the waggle dance: *boss* and *hnRNP A1*. Overall, our study demonstrates the application of ML to analyse transcriptomics data and identify candidate genes underlying social behaviour. This approach has great potential for application to a wide range of different scenarios in evolutionary ecology, when investigating the genomic basis for complex phenotypic traits, and can present some clear advantages compared to the established tools of gene expression analysis, making it a valuable complement for future studies.

KEYWORDS
bioinfomatics, feature selection, genomics, gene structure and function, insects, social evolution

1 | INTRODUCTION

The complex relationship between genes and behaviour has fuelled a large body of recent research (Robinson, 2004; Weitekamp et al., 2017) and we now know that gene activity can influence brain function, which in turn may affect behaviour (Robinson et al., 2008). Several studies have shown that behavioural states (distinct and well-characterized behaviours such as foraging or defensive behaviour) can be associated with distinct gene expression profiles in neural tissue, representing the basis for the neurogenomic approach: for example, large gene networks have been associated with foraging and defence behaviour in honeybees (Hunt et al., 2007), and numerous candidate neurological genes have been linked to aggression in a variety of organisms, including honeybees (Liu et al., 2016) and zebrafish (Filby et al., 2010). Nonetheless, most studies have focused on behavioural states that are long lasting or inherent to a species (Zayed & Robinson, 2012), whereas more plastic and transient social interactions among members of the same species (or colony) have been less characterized at the neurogenomic level (Taylor et al., 2021). This is probably due to the challenges associated with combining accurate behavioural observations with complex experimental designs to obtain and analyse large sets of gene expression data (Robinson et al., 2008).

The Western honeybee *Apis mellifera* has become a model organism for neurogenomics due to its fascinating sociobiology, the ecosystem services it provides as a pollinator and the availability of a fully annotated genome (Weinstock et al., 2006). Honeybees display perhaps one of the most iconic social behaviours in the animal world—the "waggle dance"—where foragers communicate the location of suitable food sources and possible nest locations to nestmates via stereotyped movements (Couvillon et al., 2012). This complex behaviour was described for the first time in the last century (von Frisch, 1967, 1974) and since then many details of its ecological, evolutionary and physiological underpinnings have been characterized (reviewed in Barron & Plath, 2017; Dyer, 2002; Price & Grüter, 2015). Despite this, we still do not have a complete picture of how the waggle dance is regulated at the brain level. Pioneering studies have started to reveal some of the key players at the levels of molecules (Barron et al., 2007; Kennedy et al., 2021; Linn et al., 2020), cell types (Kiya et al., 2007) and genetic pathways (Sen Sarma et al., 2009, 2010) associated with dance communication, but it is unclear what genes in the honeybee brain trigger the performance of dance behaviour once activated.

Traditionally, the neurogenomic approach has consisted of using statistical methods to calculate differential gene expression (Fang et al., 2012), which requires robust data analysis techniques due to the large volumes of sequence reads generated per sample (Kukurba & Montgomery, 2015). An interesting development in the field to address the increased computational needs of these approaches has been the application of machine learning (ML) to genomics studies (Libbrecht & Noble, 2015). ML is a branch of computer science which focuses on the study of algorithms that can improve automatically through experience or by the use of data. These algorithms were proposed to address complex problems which could not be solved through an explicit list of computational steps. Thus, ML methodologies have proved to be powerful resources and have been the focus of extensive research recently to identify the possibilities of new applications to a wide range of fields in biology and medicine (Saeys et al., 2007; Wang et al., 2016; Zhou & Tuck, 2007). Despite the abundance of studies applying ML frameworks to transcriptomic data, its use to characterize the molecular regulation of highly plastic and transient behaviours has not yet been properly explored.

In this study, we sought to identify the genes associated with the performance of dance behaviour in honeybee foragers using an ML approach. We obtained a transcriptomic data set of brain tissues (mushroom bodies) from honeybee foragers that were sampled for another study designed to underpin the molecular basis for learning distance and direction through the waggle dance (Manfredini et al *in prep.*): mushroom bodies were targeted for this study as they are the best suited brain tissue to explore high cognitive functions in insects (Menzel, 2012; Peng & Chittka, 2017), including spatial tasks (Buehlmann et al., 2020; Kamhi et al., 2020). We trained three classification algorithms on the expression levels of 15,314 transcripts that equal the total number of currently known genes of the honeybee genome, with the direct goal of classifying honeybees according to whether or not they performed a waggle dance upon their return from a foraging trip (i.e., dancers vs. nondancers). Thereafter, we unified the information obtained from the different ML approaches to identify the genes associated with these complex behavioural states, and we compared these results with more traditional analyses of gene expression based on the quantification of transcript abundance across groups (namely, a Likelihood Ratio Test [LRT] and a Generalized Linear Model [GLM]). Together, our study provides deeper insight into the molecular regulations of the waggle dance, a plastic and transient behavioural state, and promotes incorporating ML in the analysis of transcriptomic data.

2 | METHODS

2.1 | Experimental setup and initial data set

The transcriptomic data used for analysis were part of an experiment prepared to study the molecular basis for social learning of distance in honeybees through the waggle dance (Manfredini, 2021). In this experiment honeybees from four different colonies were trained to visit a feeder containing a sucrose solution (concentration = 2 M) positioned at the end of a 6-m-long tunnel (Srinivasan et al., 2000), which was used to alter the bee's perception of distance as follows: vertical stripes (with respect to the direction of flight) on the tunnel walls were used to increase the estimated flight distance, while horizontal stripes were used to decrease it (Figure 1). Honeybees were then marked at the feeder according to perceived distance (similarly to Sen Sarma et al., 2010), yielding two groups: "honeybees perceiving long distance" and "honeybees perceiving short distance."

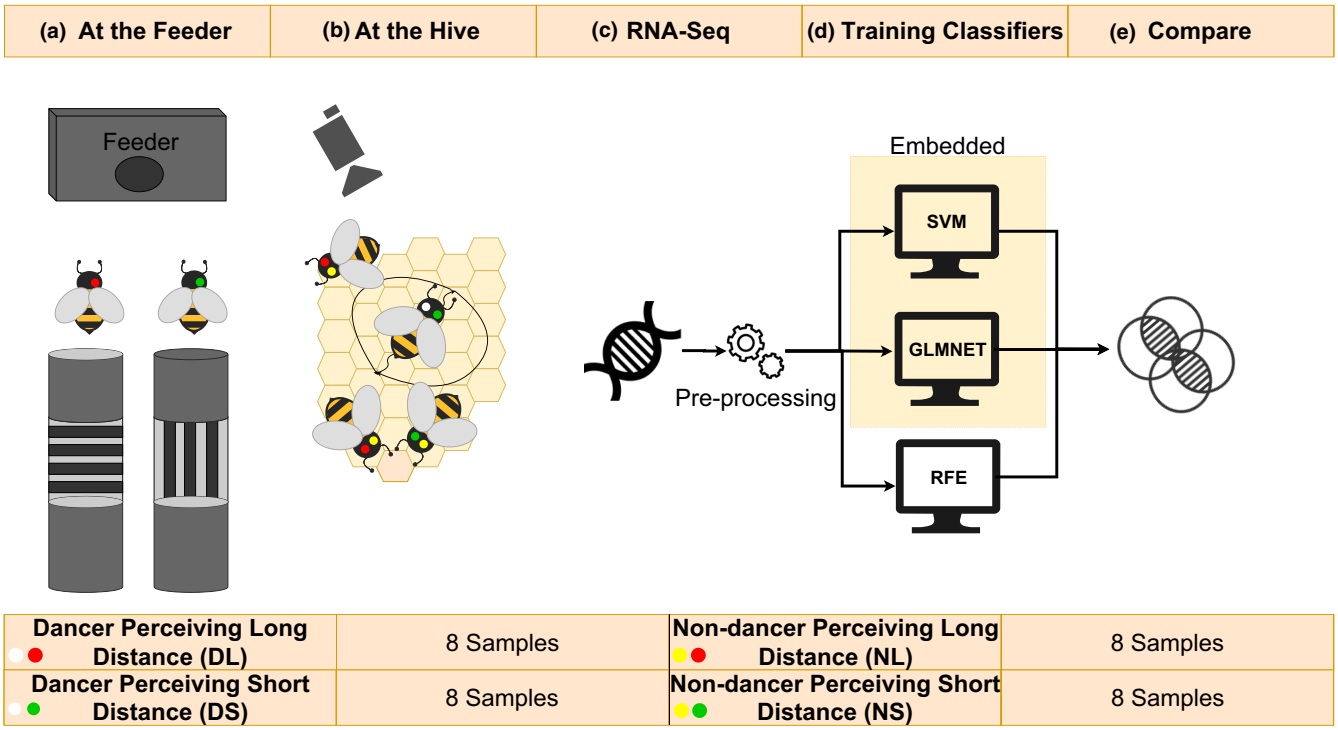

FIGURE 1 Schematic representation of the experimental design. (a) Honeybees visited a feeder through a tunnel, used to alter their distance perception (vertical stripes with respect to the direction of flight increased, while horizontal stripes decreased the distance perceived), and were then marked accordingly. (b) At the observation hive, honeybees were recorded while performing or not performing dancing behaviour and were finally prepared for RNAseq analysis (c). (d) Three machine learning algorithms (SVM, GLMNET and RFE) were trained on the preprocessed sequence reads (Training Classifiers). (e) Key features from each model were compared to identify common elements (genes or predictors)

Honeybee colonies were housed in an observation hive, which allowed direct monitoring of the comb where honeybees normally performed the waggle dance after returning from a foraging trip, known as the "dance floor" (Tautz, 1996). Bees were trained during the morning to visit the feeder at the end of the tunnel—it usually took ~5 h to complete this part—and then in the afternoon (between 2 and 4 pm) foragers that regularly visited the feeder were monitored by an observer while foragers were flying from the dance floor to the feeder and vice versa repeatedly. During this 2-h time window, the dance floor was also recorded with a video camera, producing a recording of all waggle dance events that occurred in the focal colony. At 4 pm on each day, marked bees were sampled at the feeder, by gently catching them with tweezers and quickly freezing them in liquid nitrogen. These samples were immediately housed in a −80°C freezer and stored there until we proceeded with the molecular work necessary to isolate RNA samples.

The dance floor footage was carefully analysed to identify marked honeybees that performed waggle dances upon their return from the feeder (hereafter "dancers") and separate them from those that instead were never seen performing any dance ("nondancers") for the whole duration of the 2-h videorecording, despite being visible on the dance floor upon their return from a foraging trip. We allocated to the dancers group bees that were seen performing at least one dance in the 2-h time window, but there was a certain level of variation in the group, with most bees performing multiple dances up to a maximum of 12 dances recorded for one individual. An analysis of the recorded dances confirmed that the manipulation was successful: bees exposed to vertical stripes advertised longer distances on average in their dances compared to bees exposed to horizontal stripes (Manfredini et al. *in prep*).

This resulted in the following four groups of honeybees: Dancer perceiving Long distance (DL), Dancer perceiving Short distance (DS), Nondancer perceiving Long distance (NL), and Nondancer perceiving Short distance (NS), with eight replicate samples in each of the four groups ($N = 32$). Brain tissues from all these samples were processed individually for RNA sequencing (RNAseq) analysis (see Appendix S1). Sequencing files were mapped to the most recent version of the honeybee genome (Amel_4.5) using the intron-aware STAR aligner, version 2.6.1a (Dobin et al., 2013), with default settings and using annotation information available for *Apis mellifera* on NCBI. Read counts were extracted using the *featureCounts* function from the Bioconductor R package SUBREAD version 1.8.0 (Liao et al., 2019) and following recommended parameters. The final data set, which represents the starting material for this study, included the read counts for 15,314 genes, corresponding to the whole honeybee genome across 32 bees. As we noticed some variation in library size for some of the bee samples (Dataset S3), we normalized read counts by the total library sizes to correct for the effect of possible outliers. To test for the existence of a colony effect, we used the *duplicateCorrelation* function in the R Bioconductor package LIMMA (version 3.50.0).

2.2 | Model hyperparameters and data preprocessing

We used the CARET package version 6.0-90 (Kuhn, 2008) in the programming language R version 4.1.1 (R Core Team, 2018) to train and assess the performance of the classifiers. To evaluate the models, a randomly allocated 20% of the data (six samples: DL1, DL3, DL7, NL1, NL6, NL7) were retained for testing and only the remaining 80% (26 samples) were used to train each model. Considering that the distance component had no visible effect on the distribution of our data as revealed by the initial analyses, we decided not to control the train/test split regarding the distance component. Furthermore, as part of the preprocessing requirements, the data were centred, scaled and freed from variables of (near) zero variance, to improve computation time by faster convergence. For this, the *nearZeroVar* function from the CARET package was used with the default cut-off values of 95/5 for frequency, and 10 for uniqueness. Thus, a variable was flagged if its frequency ratio (frequency of the most common value to second most common) was more than 19, and the percentage of unique values (number of unique values divided by total number of samples × 100) was below 10.

While training, we assessed the performance of each classifier on a validation set using repeated *k*-fold cross-validation (cv) on the 26 samples with 100 repeats per model (Beleites & Salzer, 2008). We chose the number of folds to be 10, a standard practice in ML (Kuhn & Johnson, 2013), which meant subdividing the training set (26 samples) into 10 bins and assessing the model performance on each one of the bins, after being retrained on the remaining nine. For each of these 10 runs, the area under the receiver operating characteristic curve (AUROC) (Marzban, 2004) was accessed using *twoClassSummary* as the summary function in the train control, and their performance was averaged to give one value for that run. Since we only had a handful of samples, the performance of the models could be highly dependent on the cv splits. Performing *k*-fold cv repeatedly (100 times) eliminated this possibility and ensured that each model was trained and validated on most (if not all) of the 26 samples.

As cv is performed to find the best parameters for the model, these 100 repeats were executed for each set of hyperparameters. The optimal hyperparameters were found by CARET implicitly, by performing a grid search through the 10 most likely values for each parameter, which were then reported.

2.3 | Selected machine learning algorithms

We used principal component analysis (PCA) (Jolliffe & Cadima, 2016) to explore the underlying structure of our data set. As a result of this set of preliminary analyses, we carefully selected the classification algorithms shown in Table 1. For a brief description of these algorithms see the Appendix S1. We also made use of "Feature Selection" techniques (FS) (Saeys et al., 2007; Wang et al., 2016) to identify the most suitable features (genes) at predicting the correlation between gene expression data and dance behaviour.

We explored three fundamentally different approaches with implicit feature ranking procedures based on previous studies (see Table 1 and also the Appendix S1): Random Forests (RF), Lasso and Elastic net Regularized Generalized Linear Model (GLMNET), and Support Vector Machine (SVM). Due to the complexity of the data, we decided to use a radial kernel for SVM, as supported by previous research (Kasnavi et al., 2018). These methods, also known as "embedded techniques," rank the features based on the already trained classifier, and as a result, the predictive power of the selected features is dependent on the performance of the model. The selected approaches proved to converge on the same final set of predictors even when subjected to repeated random starting conditions.

Whereas embedded methods obtain the importance of certain features from the trained model, wrapper methods, such as Recursive Feature Elimination (RFE), add an extra layer to the training process, and embed the model hypothesis search within the feature subset search (Saeys et al., 2007). More specifically, RFE uses backwards selection to assess the importance of each feature to the model, and discards or keeps them accordingly at each iteration. The best performing subset of features is then reported and the model is refitted on them. The ranking of the features is done by the underlying algorithm, which can be RF, SVM, GLMNET or others (Granitto et al., 2006; Li et al., 2015; Zhou & Tuck, 2007). Considering the promising properties of RF for genomic studies (Statnikov et al., 2008), we decided to use RF as the underlying model for recursive feature elimination. Even though RF had been explored as an FS algorithm in the early stages of this study, we decided to only include results of RF being run as part of the RFE procedure (RFE-RF), to avoid any overrepresentation of predictors selected by RF in the final set of focal genes.

2.4 | Characterization of focal genes

The results of the described approaches were used comparatively to characterize the relevance of the ML models and obtain a final set of predictors. First, we reported the subset of features identified by RFE-RF and compared them with the top ranked features of SVM and GLMNET, in order to contrast the three approaches, and distil an initial list of common features.

As our goal was to identify the most promising set of candidate genes and discuss them in detail, we then focused on a restricted subset of the main output. Namely, we obtained the top 20 most important features according to the individual ranking of each approach, which were then compiled into a single list of focal genes. To test the statistical significance of the overlaps, we calculated the Jaccard Index and Odds Ratio with the GeneOverlap R package (Shen, 2021). The annotations of overlapping GENES were obtained using NCBI (https://www.ncbi.nlm.nih.gov/). Where NCBI could not provide any information on putative gene function, we used BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) using default settings and the nucleotide-to-nucleotide function. We initially compared the sequence of the transcript of interest against the honeybee genome (*A. mellifera*) and then, if this did not provide any meaningful results,

TABLE 1 Benchmark algorithms. We chose to test SVM, GLMNET, RF and RFE for our study, based on their use in previous research

Algorithm	FS Method	Reviewed in:	Featured in:
Support Vector Machine (SVM)	Embedded	Noble (2006)	Aruna and Rajagopalan (2011), Guyon et al. (2002), Huang et al. (2018), Taylor et al. (2021)
Random Forest (RF)	Embedded	Breiman (2001)	Chen and Ishwaran (2012), Díaz-Uriarte and Alvarez de Andrés (2006)
Generalized Linear Model (GLMNET)	Embedded	Friedman et al. (2010)	Engebretsen and Bohlin (2019)
Recursive Feature Elimination (RFE)	Wrapper	Granitto et al. (2006), Saeys et al. (2007)	Zhou and Tuck (2007)

Note: The first three algorithms use embedded feature selection (FS) to obtain key predictors from the trained model (Embedded), while RFE requires an underlying embedded approach for the ranking (Wrapper). We report the studies that featured or reviewed these algorithms.

we compared it against the whole repository, to see whether any significant sequence similarity was detected against orthologues in other organisms (i.e., matches with high score and low E-value). We then performed overlap analyses to detect candidate genes that were in common among the three algorithms. For comparison with standard analytical methods, we also analysed the same data set of RNAseq read counts with a traditional transcriptomic approach to identify differentially expressed genes across groups (see Fang et al., 2012; Kukurba & Montgomery, 2015). We used two different statistical analyses using the Bioconductor R package: we performed LRT using DESEQ2, version 1.24 (Love et al., 2014), where we created a reduced model to test for the effect of any treatment (or behavioural group in this case, i.e., DL, DS, NL and NS) on gene expression, and we fit a simple GLM using EDGER (Robinson et al., 2010), where we grouped bees according to presence/absence of dance behaviour and we contrasted against each other ([DL + DS] vs. [NL + NS]). We followed recommended settings for both analyses (normalization performed with the variance stabilizing transformation or *vst* function in DESEQ2, and with Trimmed Mean of M-values or TMM in EDGER) and we adopted a false-discovery rate (FDR) equal to 0.05 to invoke a statistically significant difference in gene expression. Last, we compared the outputs of these analyses with the list of candidate genes from the ML approaches to identify common genes. The LRT was also used to check for the possibility of any effects due to colony of origin or lane of the sequencer used that could be responsible for driving the observed patterns of gene expression: none of these factors was associated with a significant effect (FDR > 0.05).

3 | RESULTS

3.1 | Exploratory analysis

The consensus correlation for the colony effect yielded a negative value (−0.3327), disproving that some colonies were more correlated with the dance behaviour than others, and suggesting therefore the absence of a colony effect in the data. Moreover, PCA was unable to clearly separate the four groups of bees according to the combination of dance behaviour (dancer (D)/nondancer (N)) and distance perceived (long (L)/short (S)) (Figure 2), or according to the colony of origin. However, when considering the dance factor alone, we obtained a low-dimensional representation/projection of the data using only few principal components (PCs), which produced easily distinguishable clusters of samples, that is dancers and nondancers (Figure 3). The representation was dominated by PC1, accounting for 51% of the variance in the data, while PC2 only accounted for 14.4% (Figure 3; Figure S1).

PCA clearly showed that dancers were clustered together towards the centre of the plot, showing lower variance than nondancers; this indicates more consistent global patterns of gene expression in dancers vs. nondancers. We also found four nondancers (NS2, NS4, NL6, NL5) which formed a separate cluster further along the PC1. These samples showed the highest loadings for PC1 (Dataset S1), with levels around 200, much higher than dancers (centred around 0) and the other nondancers (all below 0). We identified the three genes with maximal loadings for PC1: GB52651 (*diphthine-ammonia ligase*), GB49108 (*PDZ domain-containing protein 8*) and GB44753 (uncharacterized gene). Comparing the maximally loaded genes for PCs 1–3 (top 5,000, for consistency) and the genes identified by the embedded methods as key predictors, we found overlaps of 0% (PC1), 67.54% (PC2) and 85.39% (PC3).

Dancers showed the highest levels of positive correlation between global patterns of gene expression when represented with components PC1 and PC2 (DL = 0.711 and DS = 0.574) and showed the highest level of negative correlation on PC2 and PC3 (DL = −0.434 and DS = −0.934, Figure 2). Overall, the analysis showed a clear underlying structure in the data set with respect to the dance component (dancers vs. nondancers), while no evident structure appeared to be associated with the perceived distance (long vs. short). Based on these findings, we proceeded in our ML analyses focusing on the "dance" factor alone.

3.2 | Recursive feature elimination

The model achieved high accuracy even when using only a portion of the available features (genes in this case). The algorithm found the optimum using 5,000 of the original features (Figure S3, Dataset S2—Sheet 5), which is around 34% of the available data. The model achieved 0.99025 AUROC, 0.9615 sensitivity and 0.906 specificity on the training data. However, the model achieved similar results using only a small fraction of these features: with only 20 variables it achieved 0.9752 AUROC, 0.906 sensitivity and 0.8715 specificity on the training set. Therefore,

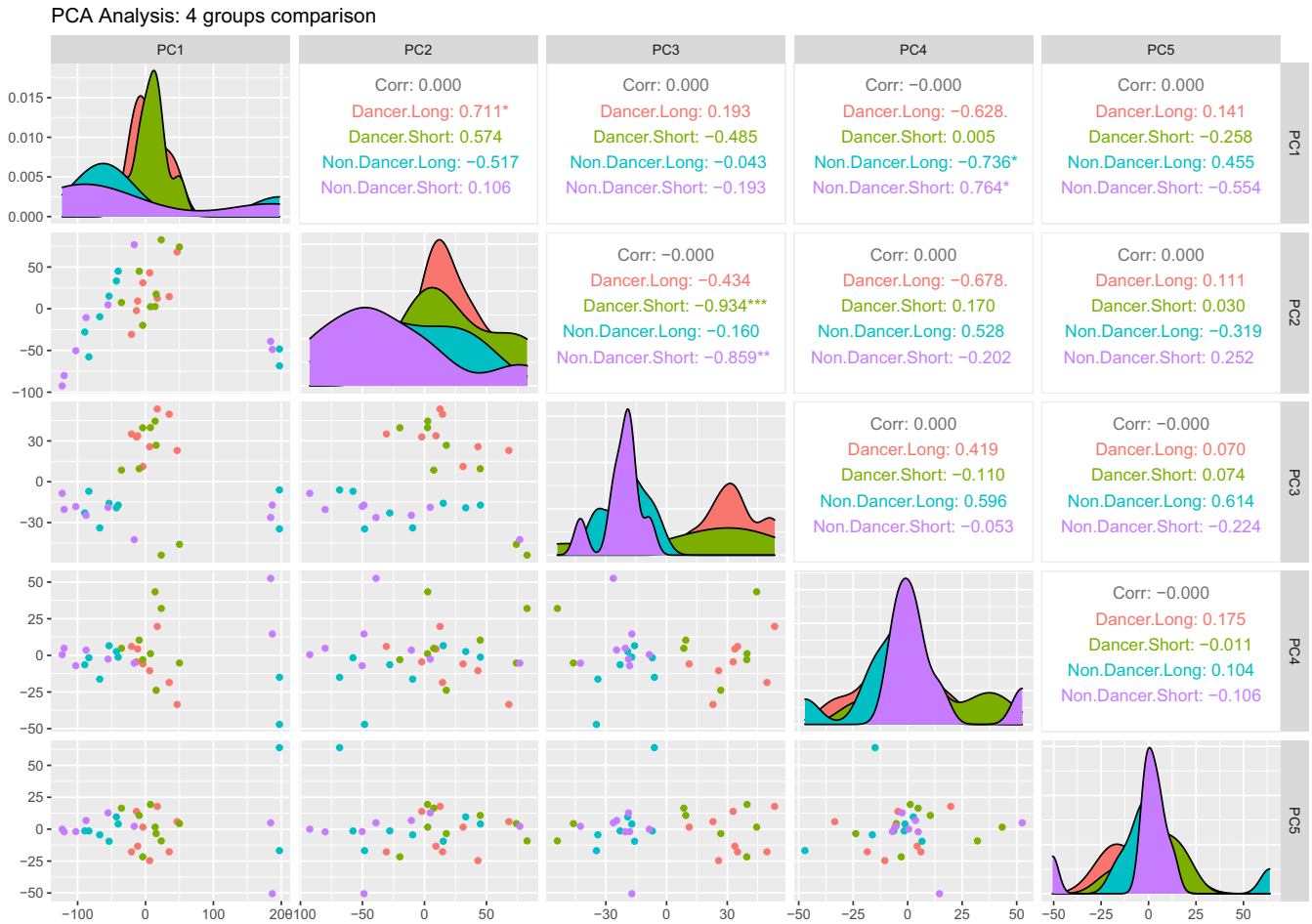

FIGURE 2 Principal component analysis. Two-dimensional comparisons of the first five principal components (PC) show clear separation between Dancers and Nondancers but not for the distance factor. The scatterplots (bottom left-hand side of the picture) show datapoints as they are represented with any two PCs. Each scatterplot corresponds to two PCs, indicated at the top of the figure and on the right: for example, the plot in the first column and fifth row corresponds to PC1 (top ID) and PC5 (right-side ID). The diagonal shows the distributions for each PC over each group in the experiment (Dancers perceiving Long distance [red], Dancers perceiving Short distance [green], Nondancers perceiving Long distance [Blue] and Nondancers perceiving Short distance [purple]). The upper right-hand side of the figure shows the correlations between each of the four groups according to the corresponding PC. These values also indicate the direction (if any) of the groupwise trends in the scatter plots. Asterisks indicate statistical significance: *$p \leq .05$, **$p \leq .01$, and ***$p \leq .001$, respectively

the model was able to represent the data using only limited information, and to generalize from the training data, obtaining 100% accuracy (ACC) on the test set. We then compared and examined the first 5,000 features (Figure S4), and then further analysed the top 20 of these genes (Figure 4), finding significant overlap with the other methods.

3.3 | Embedded methods

We trained two classifiers with the underlying algorithms SVM and GLMNET (see Table 1) using the hyperparameters as described in the Appendix S1 (Dataset S2—Sheet 1 and Sheet 3). For SVM the optimum was achieved with sigma = 4.801305e-05 and $C = 4$, with 0.99875 AUROC, 0.948 sensitivity and 0.9875 specificity. GLMNET used alpha = 0.4 and lambda = 0.03060868, to obtain 0.998 AUROC, 0.938 sensitivity and 0.9885 specificity on the training set. Both algorithms achieved 100% ACC on the test set, with a 95% likelihood that the true value laid between 54% and 100%; the wide range is due to the limited size of the test set. The No Information Rate (NIR) was 0.5, as we started from a balanced data set, and the p-value for ACC > NIR was 0.01563. We concluded that both algorithms generalized successfully, as high performance was achieved on both the training and test data sets (Figure S2).

3.4 | Overlap between selected features

The selected 5,000 features from RFE-RF (Dataset S2—Sheet 6) were compared with the top ranked genes from SVM (top 5,000 for consistency, Dataset S2—Sheet 4) and the variables selected by GLMNET (see Figures S5 and S6 for a list of the 20 most important genes for these approaches). GLMNET found only 86 genes to be important (Dataset S2—Sheet 2) and set the coefficients (importance) of the remaining 15,228 genes to 0. All these 86 genes were

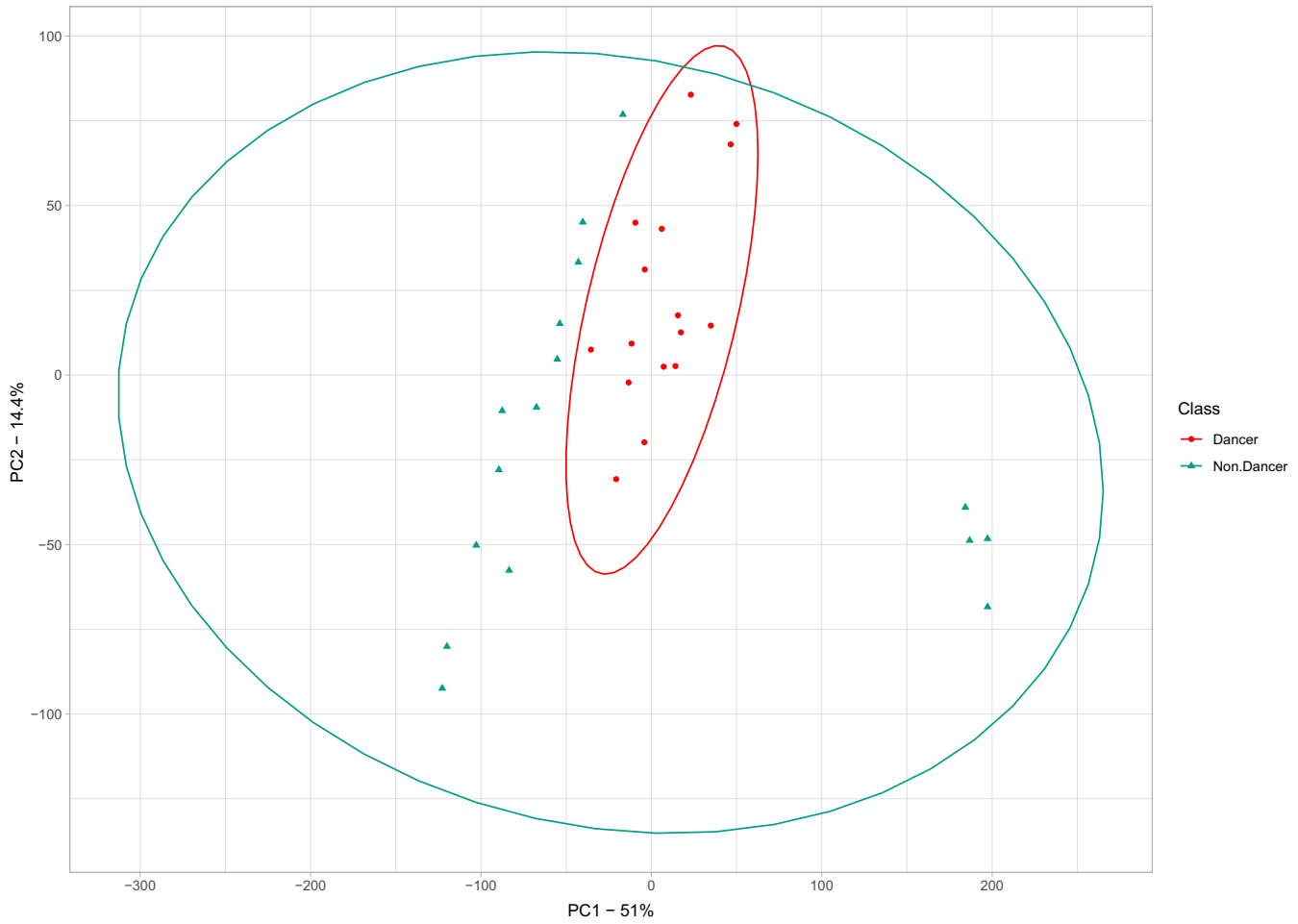

FIGURE 3 Two-dimensional projection of the data with the first two principal components. The difference between the two classes "Dancer" (red circles) and "Nondancer" (turquoise triangles) became more evident when excluding the distance factor. Ellipses show 95% confidence of the variance for the two class groups

included in the optimal subset selected by RFE-RF, with 83 of them selected by SVM as well (Jaccard Index 0.016, odds ratio 28.995, $p < .001$). The overlap between SVM and RFE-RF was also significant with 3,804 genes in common (see Figure S4), giving a Jaccard Index of 0.613, and an odds ratio of 24.246, with $p < .001$.

3.5 | Genes identified as key predictors

There were 18 genes (predictors) that were shared between at least two approaches (see Figure 4). The largest overlap was observed between RFE and SVM (16 genes) while the overlap between SVM and GLMNET was smaller (four genes). No overlap was detected between RFE and GLMNET. The Jaccard Index between SVM and RFE was the most significant (0.739), while between SVM-GLMNET and GLMNET-RFE it was 0.111 and 0.052, respectively. Similarly, the odds ratio indicated strong association between SVM and RFE (10,265.036, $p < .001$). Overall, elements in common corresponded mainly to protein coding genes, except for "GB40714," indicating noncoding RNA. We were able to retrieve functional information for most of the genes from annotations of the honeybee genome (see Table 2), with the exception of "GB50940" and" GB45448," where annotations were available only for closely related insects (*Apis dorsata* and *Apis cerana*, respectively), and "GB54617" that we could not find any information for.

3.6 | Comparison with standard gene expression analyses

We characterized gene expression patterns in the same groups of honeybees as above with standard statistical approaches to identify possible elements in common with the ML approaches that we tested. The LRT approach identified 243 genes that were statistically different between any two groups of bees, while the GLM approach identified 373 genes that were specifically different between dancers vs. nondancers (see Appendix S1 for the lists of these genes). We performed overlap analyses between these two gene sets and the list of 18 genes selected by the ML approaches. This resulted in five genes in common for the LRT approach, and nine genes in common for the GLM approach: both overlaps were statistically significant (representation factors: 17.5 and 20.5; $p < .001$ in

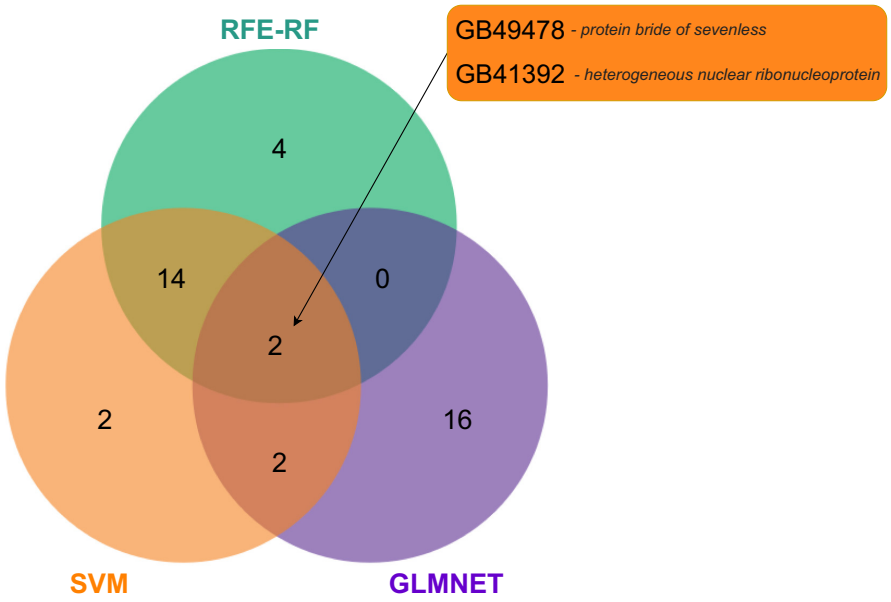

FIGURE 4 Overlap between selected features. We queried the 20 most important features in each trained classifier, SVM (orange, bottom left), GLMNET (purple, bottom right) and RFE-RF (green, top), which were then compared for overlapping subsets of genes. There were 16 genes selected both by RFE-RF and SVM, with two genes (GB41392, GB49478) selected by all three approaches. GLMNET showed little overlap with SVM and RFE-RF

both comparisons) indicating more genes in common than expected by chance. Five genes were shared across all analyses that we performed (see Table 2). Interestingly, all focal genes identified by ML approaches (as well as elements in common with gene expression analyses) were expressed at higher levels in dancers, indicating strong consistency in their expression patterns in association with the regulation of dancing behaviour (see Figure 5).

Moreover, we tested the robustness of LRT and GLM by retraining SVM and GLMNET using only the 243 and 373 identified genes. We found that the models achieved the same performance (100% ACC) on the test set given the genes identified by LRT but were inferior on the 373 genes found by GLM, as SVM achieved only 66.67% ACC, by misclassifying two dancer samples, while GLMNET achieved 100% ACC. We also compared these subsets of genes to the maximally loaded genes for PC1 (top 5,000), and we found 15.23% and 5.9% genes in common, respectively. As this overlap was absent in the case of the genes identified by the embedded methods, this led us to believe that ML methods were more robust to the separation caused by the outliers.

4 | DISCUSSION

In the present study, we implemented an ML approach to investigate the transcriptomic signatures arising from a complex plastic phenotype. We explored the unique gene expression profiles of *Apis mellifera* associated with dance behaviour in order to determine the set of focal genes that could play a key role in the regulation of this complex behaviour. Training one wrapper algorithm (RFE-RF) and two embedded models (SVM and GLMNET), we were able to achieve perfect accuracy in assigning honeybees to the major behavioural response that we tested ("dancer" vs. "nondancer") according to gene expression data. The RFE-RF approach highlighted how the genomic signature associated with the waggle dance is rather heterogeneric and can be traced across a wide portion of the honeybee genome (one third). At the same time, using FS and comparative analyses, we were able to obtain a restricted set of key predictors for each classifier, which were then distilled into a list of genes. Our results show that ML models can be used in addition to standard methods of gene expression analysis and as a complementary approach to characterize the transcriptomic profile associated with the honeybee waggle dance and to identify sets of genes that are promising candidates for the regulation of dance behaviour.

In our initial preliminary analyses (PCA) we were able to clearly separate dancers from nondancers, except for those four nondancer bees that we identified as forming a separate cluster. We exclude the possibility that these samples might represent a set of "outliers" significantly driving the outcome of our analyses. First, there is nothing visibly different associated with them: they came from different colonies (actually from all four colonies tested in our assays), were sequenced in different lanes and produced libraries of different size (which we controlled for in our normalization step). Furthermore, the lack of overlap between the best candidate genes from our ML approach and the maximum loadings for PC1 fully supports the fact that the molecular separation between dancers and nondancers was not driven by these four samples. On the other hand, the PCA approach was unable to detect any major effects of distance perception, which was one of the research questions that we had initially pursued. Although we found that the impact of distance perception on gene expression was too subtle to be detected by our approaches, other studies have succeeded in identifying the effect of distance perception alone on honeybee brain gene expression, using more traditional statistical tools of transcriptomic analyses (Sen Sarma et al., 2010). It is possible that with an increased sample size, we would have been able to investigate this behaviour further. Alternatively, the transcriptomic signature associated with distance perception might be more significant in honeybees experiencing real distance as opposed to the perceived distance that honeybees experienced through our tunnel manipulation setup. In fact, a larger set of genes was found to differ between foragers experiencing real long

TABLE 2 Focal genes for dance behaviour

Gene ID	Gene name	Gene description	SVM	GLMNET	RFE-RF	Gene expression analysis	References
GB41387	Gpdh	*Glycerol-3-phosphate dehydrogenase*	X	X		N/A	Merritt et al. (2006), Wilanowski and Gibson (1998)
GB40722	LOC102655540	*Cytochrome c oxidase assembly protein COX19*	X	X		GLM	Gaudet et al. (2011)
GB50642	LOC550745	*NEDD8-activating enzyme E1 regulatory subunit*	X		X	N/A	Liu et al. (2016)
GB40714	LOC102656214	Uncharacterized	X		X	GLM & LRT	
GB54617	NA	NA	X		X	GLM & LRT	
GB50940	LOC102681297	*Phosphatidate phosphatase LPIN3*	X		X	GLM & LRT	
GB49348	LOC409041	*Transmembrane protein 115*	X		X	N/A	
GB41887	LOC100577280	Uncharacterized	X		X	N/A	
GB41392	LOC552027	*Heterogeneous nuclear ribonucleoprotein A1, A2/B1 homologue*	X	X	X	GLM	Liu et al. (2017)
GB55807	LOC413341	*Helicase domino*	X		X	N/A	Ruhf et al. (2001)
GB50290	LOC724917	*Inactive serine protease scarface*	X		X	GLM & LRT	Srivastava and Dong (2015)
GB42813	LOC724440	*Calcium-transporting ATPase type 2C member 1*	X		X	GLM	Banerjee et al. (2006), Roti et al. (2013)
GB49478	LOC726228	*Protein bride of sevenless*	X	X	X	GLM & LRT	Bao and Friedrich (2008), Kitadate et al. (2007)
GB55460	LOC552092	Uncharacterized	X		X	GLM	
GB55243	LOC551425	*UDP - N-acetylglucosamine—peptideN-acetylglucosaminyltransferase 110-kDa subunit*	X		X	N/A	
GB42174	LOC726609	*N-alpha-acetyltransferase 11*	X		X	N/A	
GB45448	LOC107999677	Uncharacterized	X		X	N/A	
GB52425	LOC408693	*Epidermal growth factor receptor kinase substrate 8*	X		X	N/A	Tognon et al. (2014)

Note: The annotations of overlapping genes (selected by at least two approaches) were obtained using NCBI or BLAST search. Gene expression analysis (GLM and LRT) detected overlapping genes, as indicated in the relevant column. The "Reference" column indicates key studies that focused on the selected genes.

Banerjee et al. (2006); Bao and Friedrich (2008); Kitadate et al. (2007); Liu et al. (2016); Liu et al. (2017); Roti et al. (2013); Ruhf et al. (2001); Srivastava and Dong (2015); Tognon et al. (2014).

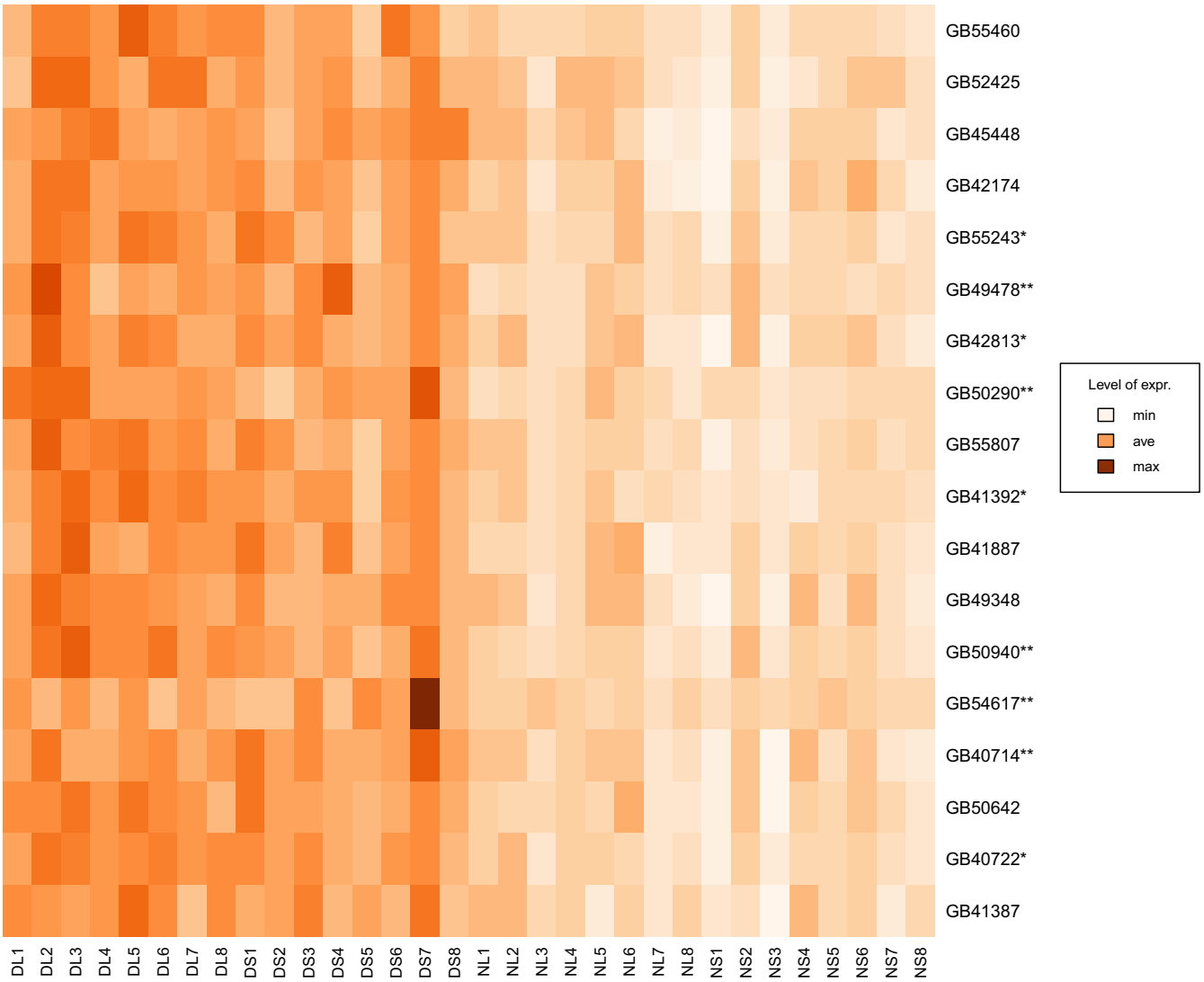

FIGURE 5 Focal genes heatmap. The heatmap shows the expression patterns for the 18 focal genes across all bee samples (Dancers grouped to the left, Nondancers to the right). Expression patterns are shown as the logarithmic transformation ($\log_{10}$) of the number of read counts for the focal genes per million counts total (or logCPM). All focal genes showed higher levels of expression in Dancers (indicated by darker colours). The gene IDs marked with a single asterisk (*) were identified by gene expression analysis (GLM approach only), while IDs marked with two asterisks (**) were identified by both GLM and LRT

distance vs. short distance (Manfredini et al., *in prep.*) but we cannot exclude that a proportion of these genes might have changed their patterns of expression from one group to the other according, for example, to different metabolic costs of flight.

It is also worth noting that GLMNET has a fundamentally different strategy towards overfitting than the other approaches, as the regularization parameter controls the cost of nonzero coefficients in the model, whereas in SVM it controls the penalty of misclassification, and RF is known not to overfit as the number of trees increases (Breiman, 2001) (see Appendix S1 for further details). However, we did not expect GLMNET to achieve perfect accuracy while discarding most of its variables (~ 99.44%), and the fact that these genes were a subset of the ones selected by other approaches is an exciting result. This also shows how a combination of different tools is the best approach for identifying candidate genes.

The extensive overlap between the three approaches, and the fact that many of the identified genes were also in common with traditional methods of transcriptomic analyses, shows great promise. We hypothesize that these genes are the best predictors for the dance behaviour, as they all appear to be expressed at higher levels in dancers vs. nondancers. In particular, the two genes that are in common to all three ML approaches and were also identified by at least one approach to gene expression analysis deserve special attention. *Boss* (*bride of sevenless*) belongs to the group of G-protein-coupled receptors, an important family of genes often associated with expression of behaviour in insects. In particular, *boss* has been linked to a set of different functions in *Drosophila*, including sight and eye development, energy homeostasis and response to glucose (Kohyama-Koganeya et al., 2015). *Boss* might have been co-opted in honeybees to regulate dance behaviour, an energetically expensive

activity that is highly related to feeding behaviour (and therefore to sugar response) and relies on visual input for orientation purposes during flight to a foraging site. In support of our findings, two previous studies addressing the regulation of the honeybee waggle dance at the molecular level have also identified genes linked to metabolism and energy production—even though *boss* is not among them (Sen Sarma et al., 2009, 2010). Interestingly, in one of these studies, genes associated with metabolism were more highly expressed in *A. mellifera* compared to a different honeybee species (*Apis florea*) that performs a simplified version of the waggle dance, further supporting the evidence that energy costs of the waggle dance can be quantified at the level of gene expression in the mushroom bodies. As for *heterogeneous nuclear ribonucleoprotein A1*, studies on the *Drosophila* orthologue *HRP59* have revealed a role for this gene in alternative splicing (Hase et al., 2006), a molecular process that allows the translation of a single mRNA molecule into multiple protein variants (Wang et al., 2015), significantly increasing the repertoire of responses to a stimulus. Even though the role of alternative splicing in the regulation of behaviour is largely unknown, this process has started to be characterized in multiple organisms, including honeybees (Foret et al., 2012), hinting at the possibility that the honeybee orthologue of *HRP59* might contribute to the high plasticity that is necessary to regulate a complex behaviour such as the waggle dance.

More functional approaches are needed to move beyond correlation and investigate whether a causal link exists between the expression levels of the genes that we identified and the performance of dance behaviour. For example, a recent study has revealed that gene expression associated with sensory perception rather than high cognitive functions is more important for bees following a dance when deciding whether to use personal information vs. social cues (the waggle dance) when engaging in the next foraging trip (Kennedy et al., 2021): it could be tested whether sensory perception has a role in the regulation of dance behaviour as well, by analysing gene expression in other brain parts such as the antennal or optic lobes that are clearly linked to the processing of sensory inputs. We are aware that other internal or external factors could contribute to define the patterns of gene expression in the honeybees that we analysed, as the brain is a complex organ that responds to a wide range of factors and stimuli that we could not control totally, such as bee age or number of dances performed, just to mention a few. However, we are confident that our strict experimental design enabled us to focus on the set of genes that are most relevant for the performance of the waggle dance *per se*: we restricted our analysis to the mushroom bodies (where transcriptomic patterns should mainly reflect behavioural responses, in particular in association with high cognitive functions) (Buehlmann et al., 2020; Kamhi et al., 2020); we sampled honeybees after they spent a whole day foraging back and forth from/to the same food source (2% sucrose solution, identical for all bees); and we analysed forager honeybees (normally from 2 to 3 weeks of age for colonies of size similar to ours; *personal observations*). If further research were to support our findings, these results could then be used to test the recruitment potential in a specific colony. By designing a diagnostic tool to directly measure the levels of expression of the focal genes and compare them against a reference, it would be possible to assess the overall ability of a colony at recruiting to a foraging site through dancing.

In conclusion, with this study we provide support for using ML models as a complementary approach to standard gene expression analyses to understand the molecular regulation of a behavioural phenotype. We show the potential of ML models to represent complex patterns in a high-dimensional data set with limited information (only 0.56% of the honeybee genome in the case of GLMNET), and the unique ability of ML approaches to generalize transcriptomic patterns from a training set of gene expression data, which testify to the predictive power of these tools (Fountain-Jones et al., 2021; Smith et al., 2020). This ability to capture transient and nonlinear responses with very little statistical supervision, and to treat each feature (or gene) as a dynamic component of a larger and cohesive picture (the phenotype of interest) is fundamentally different from other tools that are more frequently used to analyse transcriptomic data, such as DESEQ2 or EDGER. These tools are based on the straightforward quantification of transcript abundance for individual genes and treat each gene as a separate entity, using stringent parameters such as fold-changes and *p*-values to decide what genes (among those showing any difference in expression levels) are biologically relevant for the patterns observed. Furthermore, the ML models that we propose here appear to be particularly robust to the presence of few samples that might "behave" differently compared to other members of the same cohort and do not fully fit within the characteristics of their experimental group (something that can seriously limit the power of standard analyses of gene expression, in particular when the sample size is small). This is evident from the lack of overlap between the features selected by the ML models and the genes with maximum loading for PC1 in the PCA, representing the genes most significantly associated with the four nondancer bees that formed a separate cluster. With high probability, the most striking feature of ML models as applied to transcriptomic analyses is their ability to classify samples of unknown phenotype—including those showing intermediate features (Taylor et al., 2021)—according to a range of parameters of interest, so that additional individuals can be added to a study in future analyses. In the context of the honeybee waggle dance, this feature could be exploited, for example, to understand what other factors influence the performance of the dance behaviour (e.g., age of the bee, colony of origin or foraging patterns) and without the need to carefully monitor every single dance event occurring in the hive, which can seriously limit any experimental design.

ACKNOWLEDGEMENTS

We thank Dr Georgios Leontidis (The School of Natural and Computing Science, University of Aberdeen) for his valuable support during the selection and implementation of the ML models, and the two anonymous reviewers for providing useful feedback that helped improve the clarity and soundness of the manuscript. We

are also grateful to NERC (Natural Environment Research Council) for funding this project and supporting MV's salary over 10 weeks through their Research Experience Placement programme (DTG reference: NE/S007377/1). The honeybee work that was performed to obtain the sequencing data used in this study was funded by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant no. 638873 to EL). This funding also supported FM during the execution of the field and molecular work.

CONFLICT OF INTEREST

The author declare no conflict of interests.

AUTHOR CONTRIBUTIONS

Designed research: MV, JM, EL, FM. Collected samples and prepared them for analysis: FM. Performed research: MV. Contributed new reagents and analytical tools: EL. Analysed data: MV, FM. Wrote the paper: MV with input from JM, EL and FM.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://github.com/Vejni/WaggleDance_MachineLearning.

DATA AVAILABILITY STATEMENT

All codes used in the analyses here reported are visible in a GitHub repository associated with this project: https://github.com/Vejni/WaggleDance_MachineLearning. The raw sequencing data that represent the starting material for the analyses here described have been deposited on NCBI SRA (Bioproject PRJNA756776).

ORCID

Marcell Veiner https://orcid.org/0000-0003-4179-1374
Juliano Morimoto https://orcid.org/0000-0003-3561-1920
Ellouise Leadbeater https://orcid.org/0000-0002-4029-7254
Fabio Manfredini https://orcid.org/0000-0002-9134-3994

REFERENCES

Aruna, S., & Rajagopalan, S. (2011). A Novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. *International Journal of Computer Applications*, *31*(8).

Banerjee, S., Joshi, R., Venkiteswaran, G., Agrawal, N., Srikanth, S., Alam, F., & Hasan, G. (2006). Compensation of Inositol 1,4,5-Trisphosphate Receptor Function by Altering Sarco-Endoplasmic Reticulum Calcium ATPase Activity in the Drosophila Flight Circuit. *Journal of Neuroscience*, *26*(32), 8278–8288. https://doi.org/10.1523/JNEUROSCI.1231-06.2006

Bao, R., & Friedrich, M. (2008). Fast co-evolution of sevenless and bride of sevenless in endopterygote insects. *Development Genes and Evolution*, *218*(3-4), 215–220. https://doi.org/10.1007/s00427-007-0201-0

Barron, A. B., Maleszka, R., Meer, R. K., & Robinson, G. E. (2007). Octopamine modulates honey bee dance behavior. *Proceedings of the National Academy of Sciences*, *104*(5), 1703–1707. https://doi.org/10.1073/pnas.0610506104

Barron, A. B., & Plath, J. A. (2017). The evolution of honey bee dance communication: a mechanistic perspective. *Journal of Experimental Biology*, *220*(23), 4339–4346. https://doi.org/10.1242/jeb.142778

Beleites, C., & Salzer, R. (2008). Assessing and improving the stability of chemometric models in small sample size situations. *Analytical and Bioanalytical Chemistry*, *390*(5), 1261–1271. https://doi.org/10.1007/s00216-007-1818-6

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Buehlmann, C., Wozniak, B., Goulard, R., Webb, B., Graham, P., & Niven, J. E. (2020). Mushroom bodies are required for learned visual navigation, but not for innate visual behavior, in ants. *Current Biology*, *30*(17), 3438–3443. https://doi.org/10.1016/j.cub.2020.07.013

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, *99*(6), 323–329. https://doi.org/10.1016/j.ygeno.2012.04.003

Couvillon, M. J., Riddell Pearce, F. C., Harris-Jones, E. L., Kuepfer, A. M., Mackenzie-Smith, S. J., Rozario, L. A., Schürch, R., & Ratnieks, F. L. W. (2012). Intra-dance variation among waggle runs and the design of efficient protocols for honey bee dance decoding. *Biology Open*, *1*(5), 467–472. https://doi.org/10.1242/bio.20121099

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*(1), 1-13. https://doi.org/10.1186/1471-2105-7-3

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dyer, F. C. (2002). The biology of the dance language. *Annual Review of Entomology*, *47*(1), 917–949. https://doi.org/10.1146/annurev.ento.47.091201.145306

Engebretsen, S., & Bohlin, J. (2019). Statistical predictions with glmnet. *Clinical Epigenetics*, *11*(1), 1–3. https://doi.org/10.1186/s13148-019-0730-1

Fang, Z., Martin, J., & Wang, Z. (2012). Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. *Cell & Bioscience*, *2*(1), 26. https://doi.org/10.1186/2045-3701-2-26

Filby, A. L., Paull, G. C., Hickmore, T. F., & Tyler, C. R. (2010). Unravelling the neurophysiological basis of aggression in a fish model. *BMC Genomics*, *11*(1), 498. https://doi.org/10.1186/1471-2164-11-498

Foret, S., Kucharski, R., Pellegrini, M., Feng, S., Jacobsen, S. E., Robinson, G. E., & Maleszka, R. (2012). DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. *Proceedings of the National Academy of Sciences*, *109*(13), 4968–4973. https://doi.org/10.1073/pnas.1202392109

Fountain-Jones, N. M., Smith, M. L., & Austerlitz, F. (2021). Machine learning in molecular ecology. *Molecular Ecology Resources*, *21*(8), 2589–2597.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1. https://doi.org/10.18637/jss.v033.i01

Gaudet, P., Livstone, M. S., Lewis, S. E., & Thomas, P. D. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics*, *12*(5), 449–462. https://doi.org/10.1093/bib/bbr042

Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, *83*(2), 83–90. https://doi.org/10.1016/j.chemolab.2006.01.007

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, *46*(1/3), 389–422. https://doi.org/10.1023/A:1012487302797

Hase, M. E., Yalamanchili, P., & Visa, N. (2006). The drosophila heterogeneous nuclear ribonucleoprotein M Protein, HRP59, regulates alternative splicing and controls the production of its own mRNA. *Journal of Biological Chemistry*, *281*(51), 39135–39141. https://doi.org/10.1074/jbc.M604235200

Huang, C., Clayton, E. A., Matyunina, L. V., McDonald, L. D. E., Benigno, B. B., Vannberg, F., & McDonald, J. F. (2018). Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. *Scientific Reports*, *8*(1), 1–8. https://doi.org/10.1038/s41598-018-34753-5

Hunt, G. J., Amdam, G. V., Schlipalius, D., Emore, C., Sardesai, N., Williams, C. E., Rueppell, O., Guzmán-Novoa, E., Arechavaleta-Velasco, M., Chandra, S., Fondrk, M. K., Beye, M., & Page, R. E. (2007). Behavioral genomics of honeybee foraging and nest defense. *Naturwissenschaften*, *94*(4), 247–267. https://doi.org/10.1007/s00114-006-0183-1

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kamhi, J. F., Barron, A. B., & Narendra, A. (2020). Vertical lobes of the mushroom bodies are essential for view-based navigation in Australian Myrmecia ants. *Current Biology*, *30*(17), 3432–3437. https://doi.org/10.1016/j.cub.2020.06.030

Kasnavi, S. A., Aminafshar, M., Shariati, M. M., Kashan, E. J., & Honarvar, M. (2018). The effect of kernel selection on genome wide prediction of discrete traits by Support Vector Machine. *Gene Reports*, *11*, 279–282. https://doi.org/10.1016/j.genrep.2018.04.006

Kennedy, A., Peng, T., Glaser, S. M., Linn, M., Foitzik, S., & Grüter, C. (2021). Use of waggle dance information in honey bees is linked to gene expression in the antennae, but not in the brain. *Molecular Ecology*, *30*(11), 2676–2688. https://doi.org/10.1111/mec.15893

Kitadate, Y., Shigenobu, S., Arita, K., & Kobayashi, S. (2007). Boss/Sev Signaling from Germline to Soma Restricts Germline-Stem-Cell-Niche Formation in the Anterior Region of Drosophila Male Gonads. *Developmental Cell*, *13*(1), 151–159. https://doi.org/10.1016/j.devcel.2007.05.001

Kiya, T., Kunieda, T., & Kubo, T. (2007). Increased neural activity of a mushroom body neuron subtype in the brains of forager honeybees. *PLoS One*, *2*(4), e371. https://doi.org/10.1371/journal.pone.0000371

Kohyama-Koganeya, A., Kurosawa, M., & Hirabayashi, Y. (2015). Differential effects of tissue-specific deletion of BOSS on feeding behaviors and energy metabolism. *PLoS One*, *10*(7), e0133083. https://doi.org/10.1371/journal.pone.0133083

Kuhn, M. (2008). Building predictive models in *R* using the **caret** package. *Journal of Statistical Software*, *28*(5), 1-26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*, Vol. *26*. Springer.

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), pdb.top084970. https://doi.org/10.1101/pdb.top084970

Li, X., Liu, T., Tao, P., Wang, C., & Chen, L. (2015). A highly accurate protein structural class prediction approach using auto cross covariance transformation and recursive feature elimination. *Computational Biology and Chemistry*, *59*, 95–100. https://doi.org/10.1016/j.compbiolchem.2015.08.012

Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, *47*(8), e47. https://doi.org/10.1093/nar/gkz114

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332. https://doi.org/10.1038/nrg3920

Linn, M., Glaser, S. M., Peng, T., & Grüter, C. (2020). Octopamine and dopamine mediate waggle dance following and information use in honeybees. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1936), 20201950. https://doi.org/10.1098/rspb.2020.1950

Liu, H., Robinson, G. E., & Jakobsson, E. (2016). Conservation in mammals of genes associated with aggression-related behavioral phenotypes in honey bees. *PLOS Computational Biology*, *12*(6), e1004921. https://doi.org/10.1371/journal.pcbi.1004921

Liu, T.-Y., Chen, Y.-C., Jong, Y.-J., Tsai, H.-J., Lee, C.-C., Chang, Y.-S., Chang, J.-G., & Chang, Y.-F. (2017). Muscle developmental defects in heterogeneous nuclear Ribonucleoprotein A1 knockout mice. *Open Biology*, *7*(1), 160303. https://doi.org/10.1098/rsob.160303

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/s13059-014-0550-8

Manfredini, F. (2021). *Machine learning and honeybee waggle dance*. NCBI SRA Bioproject PRJNA756776.

Marzban, C. (2004). The ROC curve and the area under it as performance measures. *Weather and Forecasting*, *19*(6), 1106–1114. https://doi.org/10.1175/825.1

Menzel, R. (2012). The honeybee as a model for understanding the basis of cognition. *Nature Reviews Neuroscience*, *13*(11), 758–768. https://doi.org/10.1038/nrn3357

Merritt, T. J. S., Sezgin, E., Zhu, C.-T., & Eanes, W. F. (2006). Triglyceride Pools, Flight and Activity Variation at the *Gpdh* Locus in *Drosophila melanogaster*. *Genetics*, *172*(1), 293–304. https://doi.org/10.1534/genetics.105.047035

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Peng, F., & Chittka, L. (2017). A simple computational model of the bee mushroom body can explain seemingly complex forms of olfactory learning and memory. *Current Biology*, *27*(2), 224–230. https://doi.org/10.1016/j.cub.2016.10.054

Price, R. I., & Grüter, C. (2015). Why, when and where did honey bee dance communication evolve? *Frontiers in Ecology and Evolution*, *3*, 125. https://doi.org/10.3389/fevo.2015.00125

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Robinson, G. E. (2004). GENOMICS: Beyond Nature and Nurture. *Science*, *304*(5669), 397–399. https://doi.org/10.1126/science.1095766

Robinson, G. E., Fernald, R. D., & Clayton, D. F. (2008). Genes and social behavior. *Science*, *322*(5903), 896–900.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Roti, G., Carlton, A., Ross, K. N., Markstein, M., Pajcini, K., Su, A. H., Perrimon, N., Pear, W. S., Kung, A. L., Blacklow, S. C., Aster, J. C., & Stegmaier, K. (2013). Complementary genomic screens identify SERCA as a therapeutic target in NOTCH1 mutated cancer. *Cancer Cell*, *23*(3), 390–405. https://doi.org/10.1016/j.ccr.2013.01.015

Ruhf, M. L., Braun, A., Papoulas, O., Tamkun, J. W., Randsholt, N., & Meister, M. (2001). The domino gene of Drosophila encodes novel members of the SWI2/SNF2 family of DNA-dependent ATPases, which contribute to the silencing of homeotic genes. *Development*, *128*(8), 1429–1441. https://doi.org/10.1242/dev.128.8.1429

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

Sen Sarma, M., Rodriguez-Zas, S. L., Gernat, T., Nguyen, T., Newman, T., & Robinson, G. E. (2010). Distance-responsive genes found in dancing honey bees. *Genes, Brain and Behavior*, *9*(7), 825–830. https://doi.org/10.1111/j.1601-183X.2010.00622.x

Sen Sarma, M., Rodriguez-Zas, S. L., Hong, F., Zhong, S., & Robinson, G. E. (2009). Transcriptomic profiling of central nervous system regions in three species of honey bee during dance communication behavior. *PLoS One*, *4*(7), e6408. https://doi.org/10.1371/journal.pone.0006408

Shen, L. (2021). *GeneOverlap: An R package to test and visualize gene overlaps*. http://shenlab-sinai.github.io/shenlab-sinai/

Smith, A. M., Walsh, J. R., Long, J., Davis, C. B., Henstock, P., Hodge, M. R., Maciejewski, M., Mu, X. J., Ra, S., Zhao, S., Ziemek, D., & Fisher, C. K. (2020). Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC Bioinformatics*, *21*(1), 1–18. https://doi.org/10.1186/s12859-020-3427-8

Srinivasan, M. V., Zhang, S., Altwein, M., & Tautz, Jürgen (2000). Honeybee navigation: nature and calibration of the "Odometer". *Science*, *287*(5454), 851–853. https://doi.org/10.1126/science.287.5454.851

Srivastava, A., & Dong, Q. (2015). Regulation of a serine protease homolog by the JNK pathway during thoracic development of *Drosophila melanogaster*. *FEBS Open Bio*, *5*(1), 117–123. https://doi.org/10.1016/j.fob.2015.01.008

Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, *9*(1), 1–10. https://doi.org/10.1186/1471-2105-9-319

Tautz, J. (1996). Honeybee waggle dance: recruitment success depends on the dance floor. *Journal of Experimental Biology*, *199*(6), 1375–1381. https://doi.org/10.1242/jeb.199.6.1375

Taylor, B. A., Cini, A., Wyatt, C. D. R., Reuter, M., & Sumner, S. (2021). The molecular basis of socially mediated phenotypic plasticity in a eusocial paper wasp. *Nature Communications*, *12*(1), 1–10. https://doi.org/10.1038/s41467-021-21095-6

Tognon, E., Wollscheid, N., Cortese, K., Tacchetti, C., & Vaccari, T. (2014). ESCRT-0 is not required for ectopic notch activation and tumor suppression in drosophila. *PLoS One*, *9*(4), e93987. https://doi.org/10.1371/journal.pone.0093987

von Frisch, K. (1967). *The dance language and orientation of bees*. Harvard University Press.

von Frisch, K. (1974). Decoding the language of the bee. *Science*, *185*(4152), 663–668.

Wang, L., Wang, Y., & Chang, Q. (2016). Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods*, *111*, 21–31. https://doi.org/10.1016/j.ymeth.2016.08.014

Wang, Y., Liu, J., Huang, B. O., Xu, Y.-M., Li, J., Huang, L.-F., Lin, J., Zhang, J., Min, Q.-H., Yang, W.-M., & Wang, X.-Z. (2015). Mechanism of alternative splicing and its regulation. *Biomedical Reports*, *3*(2), 152–158. https://doi.org/10.3892/br.2014.407

Weinstock, G. M., Robinson, G. E., Gibbs, R. A., Weinstock, G. M., Weinstock, G. M., & Robinson, G. E. (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, *443*(7114), 931–949. https://doi.org/10.1038/nature05260

Weitekamp, C. A., Libbrecht, R., & Keller, L. (2017). Genetics and evolution of social behavior in insects. *Annual Review of Genetics*, *51*(1), 219–239. https://doi.org/10.1146/annurev-genet-120116-024515

Wilanowski, T. M., & Gibson, J. B. (1998). sn-Glycerol-3-phosphate dehydrogenase in the honey bee *Apis mellifera*—an unusual phenotype associated with the loss of introns. *Gene*, *209*(1-2), 71–76. https://doi.org/10.1016/S0378-1119(98)00016-X

Zayed, A., & Robinson, G. E. (2012). Understanding the relationship between brain gene expression and social behavior: lessons from the honey bee. *Annual review of genetics*, *46*, 591–615. https://doi.org/10.1146/annurev-genet-110711-155517

Zhou, X., & Tuck, D. P. (2007). MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics*, *23*(9), 1106–1114. https://doi.org/10.1093/bioinformatics/btm036

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Veiner, M., Morimoto, J., Leadbeater, E., & Manfredini, F. (2022). Machine learning models identify gene predictors of waggle dance behaviour in honeybees. *Molecular Ecology Resources*, *22*, 2248–2261. https://doi.org/10.1111/1755-0998.13611